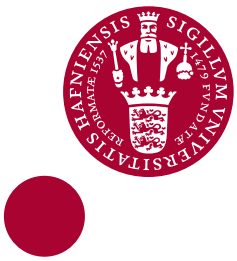

Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles

Prieto, L; Haned, H; Mosquera, A; Crespillo, M; Alemañ, M; Aler, M; Alvarez, F; Baeza-Richer, C; Dominguez, A; Doutremepuich, C; Farfán, M J; Fenger-Grøn, Martin; García-Ganivet, J M; González-Moya, E; Hombreiro, L; Lareu, M V; Martínez-Jarreta, B; Merigioli, S; Milans Del Bosch, P; Morling, N; Muñoz-Nieto, M; Ortega-González, E; Pedrosa, S; Pérez, R; Solís, C; Yurrebaso, I; Gill, P

Published in:
Forensic science international. Genetics

DOI:
10.1016/j.fsigen.2013.10.011

Publication date:
2014

Forensic Science International: Genetics 9 (2014) 47–54

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles

L. Prieto [a], H. Haned [b], A. Mosquera [c], M. Crespillo [d], M. Alemañ [e], M. Aler [f], F. Álvarez [a], C. Baeza-Richer [g], A. Dominguez [h], C. Doutremepuich [i], M.J. Farfán [j,k], M. Fenger-Grøn [l], J.M. García-Ganivet [m], E. González-Moya [k], L. Hombreiro [n], M.V. Lareu [c], B. Martínez-Jarreta [o], S. Merigioli [p], P. Milans del Bosch [q], N. Morling [l], M. Muñoz-Nieto [d], E. Ortega-González [r], S. Pedrosa [s], R. Pérez [d], C. Solís [a], I. Yurrebaso [t], P. Gill [u,v,*]

[a] Comisaría General de Policía Científica, University Institute of Research in Forensic Sciences (IUICP), Madrid, Spain
[b] Netherlands Forensic Institute, Department of Human Biological Traces, The Hague, The Netherlands
[c] Forensic Science Institute Luis Concheiro, University of Santiago de Compostela, Spain
[d] Instituto Nacional de Toxicología y Ciencias Forenses, Servicio de Biología, Barcelona, Spain
[e] Cefegen, Madrid, Spain
[f] Legal Medicine Institute of Valencia, Spain
[g] Department of Toxicology and Health Legislation, Faculty of Medicine, Complutense University of Madrid, Spain
[h] Instituto Nacional de Toxicología y Ciencias Forenses, Servicio de Biología, Sevilla, Spain
[i] Laboratoire d'Hématologie Médico-Légale, Bordeaux, France
[j] GHEP-ISFG, Grupo de habla española y portuguesa de la International Society for Forensic Genetics, Spain
[k] Instituto Nacional de Toxicología y Ciencias Forenses, Servicio de Biología, Madrid, Spain
[l] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[m] DNA Laboratory, Spanish Forensic Police, Granada, Spain
[n] DNA Laboratory, Spanish Forensic Police, A Coruña, Spain
[o] Department of Forensic Medicine, University of Zaragoza, Spain
[p] Institute of Legal Medicine, Catholic University of Sacred Heart, School of Medicine, Rome, Italy
[q] Servicio de Criminalística, Dpto. de Biología, Guardia Civil, Spain
[r] Unitat Central del Laboratori Biològic – Cos de Mossos d'Esquadra, Cataluña, Spain
[s] Nasertic, Navarra, Spain
[t] Ertzaina DNA Laboratory, Basque Country Police, Spain
[u] Norwegian Institute of Public Health, Oslo, Norway
[v] University of Oslo, Oslo, Norway

ARTICLE INFO

Article history:
Received 6 July 2013
Received in revised form 14 October 2013
Accepted 22 October 2013

Keywords:
Forensim
LRmix
Euroforgen-NoE
Drop-out
Drop-in
Likelihood ratio

ABSTRACT

There has been very little work published on the variation of reporting practices of mixtures between laboratories, but it has been previously demonstrated that there is little consistency. This is because there is no current uniformity of practice, so different laboratories will operate using different rules. The interpretation of mixtures is not solely a matter of using some software to provide 'an answer'. An assessment of a case will usually begin with a consideration of the circumstances of a crime. Assumptions made about the numbers of contributors follow from an examination of the electropherogram(s) – and these may differ between the prosecution and the defence hypotheses. There may be a necessity to evaluate several sets of hypotheses for any given case if the circumstances are uncertain. Once the hypotheses are formulated, the mathematical analysis is complex and can only be accomplished by the use of specialist software. In order to obtain meaningful results, it is essential that scientists are trained, not only in the use of the software, but also in the methodology to understand the likelihood ratio concept that is used. The Euroforgen-NoE initiative has developed a training course that utilizes the LRmix program to carry out the calculations. This software encompasses the recommendations of the ISFG DNA commissions on mixture interpretation and is able to interpret samples that may come from two or more contributors and may also be partial profiles. Recently, eighteen different laboratories were trained in the methodology. Afterwards they were asked to independently analyze two different cases with partial mixture DNA evidence and to write a statement

* Corresponding author at: Norwegian Institute of Public Health, Oslo, Norway.
E-mail addresses: peterd.gill@gmail.com, peterd.gill@virgin.net (P. Gill).

1872-4973/$ – see front matter
http://dx.doi.org/10.1016/j.fsigen.2013.10.011

court-report. We show that by introducing a structured training programme, it is possible to demonstrate, for the first time, that a high degree of standardization, leading to uniformity of results can be achieved by participating laboratories.

1. Introduction

The theory to interpret complex mixtures (defined as multi-contributor DNA profiles that are partial) is now well established. The International Society for Forensic Genetics (ISFG) DNA commission [1] recognized likelihood ratios (LRs) to be the preferred way to associate strength of evidence of DNA mixtures. Since then a number of different methods have been described [2–7].

The software used in this collaborative exercise was LRmix [8]. This open-source solution enables the calculation of LRs of complex mixtures, taking into account the twin effects of drop-out ($Pr(D)$) and drop-in ($Pr(C)$) [9]. In addition, several PCR replicates from the same DNA extract can be analyzed simultaneously. This useful feature reduces the subjectivity involved with deriving a consensus profile and avoids the associated problems [10–12]. This has the advantage of incorporating all available information into a single LR calculation. With LRmix, any given number of known contributors and up to three unknowns (under each hypothesis) can be analyzed. The number of contributors can also differ under the prosecution and the defense hypotheses. However, it is usual for the suspect to be replaced by an unknown contributor under the defense hypothesis. The LRmix module also includes a performance test to further evaluate the obtained LRs; this is achieved by replacing persons of interest (suspects or victims) with simulated random persons unrelated to the case [8,13] – a true perpetrator will provide a high likelihood ratio, and a robust result is demonstrated if random man substitution gives a low LR result.

Although the theory behind LRmix was published many years ago [14,15], its complexity, coupled with the lack of software, delayed implementation in casework. Education and training of DNA experts and caseworkers is necessary before new methods can be introduced. A black box solution is not provided. There is a strong interaction between the expert and software, where the former decides the propositions that form the basis of the likelihood ratio test dependent upon the casework circumstances, the number of alleles in the electropherogram (epg), and the estimated number of contributors. It is possible that multiple sets of propositions may be tested. Consequently, there is no single answer. For this reason, the approach is described as 'exploratory' [13]. The DNA Commission of the ISFG, the ENFSI DNA Working Group and other scientific societies have strongly advocated the need for more education in this field.

There is almost no information in the literature to test consistency between laboratories [16,17] to interpret mixtures. The only published study by Duewer et al. [18] showed marked differences between laboratories to interpret minor contributors in mixed DNA profiles. This inconsistency was primarily caused by the inevitable subjectivity of diverse approaches that are engendered by 'binary methods' used to describe match/non-matching alleles. This subjectivity is completely avoided with probabilistic models since it is no longer necessary to define profiles in terms of match/non-match [9].

The European Forensic Genetics Network of Excellence (Euroforgen-NoE) is a new project funded by the Seventh Framework of the EU and is funded by the EU commission to carry out an active training programme.

In this context, Euroforgen-NoE and the Spanish Forensic Police co-funded a course on the "Interpretation on mixtures and complex DNA profiles" (lecturers Peter Gill and Hinda Haned) which was held in Madrid in 2012. The target group was DNA experts working in forensic genetics. The overall aim of the course was to introduce probabilistic methods to evaluate STR typing results that may include drop-out and/or drop-in, including an evaluation of complex mixtures using likelihood ratio methods. At the end of the course, lecturers and participants decided to organize an inter-laboratory exercise to ascertain whether standardization of approach could be demonstrated. The results of that exercise are now presented as the first demonstration of inter-laboratory standardization achieved via use of probabilistic methods to interpret mixtures.

2. Materials and methods

Participating laboratories were asked to evaluate the following two hypothetical forensic cases.

2.1. Exercise 1 circumstances (rape case)

A woman was raped in Madrid. The pathologist took a sample from the vagina of the victim and sent it to the DNA Laboratory. A suspect was detained following police investigations. The judge asked the scientist to determine if the suspect could have contributed to the vaginal sample from the victim.

2.2. Exercise 2 circumstances (homicide case)

The dismembered body of a woman was found in the countryside on the outskirts of a Spanish village. The police suspect that her husband (a butcher) committed the murder. Apart from other evidence, the body parts were perfectly separated from each other (indicating "professional" quartering) and the suspect also has a cut in his right hand. The police interrogated the man and he admitted that he had cut his hand with a knife. The apparently clean knife (evidence) was sent to the DNA Laboratory to be analyzed. Reference samples from the victim and the suspect were taken. The judge asked the scientist if DNA from the victim was present on the knife.

2.3. Information supplied to participants

For both cases, files (in CSV format) were provided to the participants (Supplementary material S1 and S2); tables of the victims' and suspects' profiles were provided, and a table including allelic frequencies for the population where the crimes were committed (Spain).

In Exercise 1, the file containing the evidence profile – epg with designations, was also provided to the participants in order to standardize the results and to avoid clerical errors leading to wrong transcription of the detected alleles to the tables (see Tables 1 and 2). In Exercise 2, only the epg (no CSV file included) from the crime-stain sample was provided in PDF format using 50 rfu as the limit of detection (Fig. 1). Exercise 2 was designed to be more flexible, so that the participants could determine if peak signals clearly above the baseline (although below the detection

Table 1
Exercise 1 profiles from the victim, the suspect and the unknown sample. Possible allelic drop-out highlighted in grey.

Locus	Victim	Suspect	Unknown sample
D10S1248	14–16	14–15	14–15–16
VWA	15–16	16–16	15–16
D16S539	12–14	9–10	9–10–12–14
D2S1338	20–21	16–20	16–20–21
D8S1179	15–15	11–13	11–13–15
D21S11	31.2–33.2	31–31.2	31–31.2–33.2
D18S51	12–15	12–14	12–14–15
D22S1045	11–15	15–16	11–15–16
D19S433	16–16.2	14–16	14–16–16.2
TH01	9–9.3	9.3–9.3	9–9.3
FGA	22–23	20–21	20–21–22–23
D2S441	10–14	11–11.3	10–11–11.3–14
D3S1358	14–15	15–18	14–15–18
D1S1656	11–14	12–17.3	11–12–14–17.3
D12S391	20–23	17–19	17–19–20–23
SE33	17–28.2	24.2–25.2	17–25.2–28.2

threshold) were to be included in the evaluation; i.e. they were able to decide which alleles formed the basis of the model in this exercise.

2.4. Method to construct a model and to interpret results

The procedure used to analyze DNA profiles using the LRmix package is explained in tutorials available from Forensim's website at http://forensim.r-forge.r-project.org/ and the process is also described in fine detail by Gill et al. [8]. In summary, to calculate a likelihood ratio from the raw data: the crime-stain evidence, the reference profiles, and the population allele frequencies are uploaded into LRmix; the known contributors (e.g. victim and suspect) and the number of unknown contributors are also input and an initial (first pass) LR is calculated by using the default parameters of $Pr(D)$, $Pr(C)$ and a term to adjust for population sub-structuring, Fst [19]. Then, a sensitivity analysis is performed. The results of this analysis consists of a two-dimensional plot of the $Pr(D)$ vs. LR (on a log 10 scale). LRmix also calculates the plausible ranges of $Pr(D)$, based on the range of results between 5 and 95 percentiles of $Pr(D)$ using a qualitative estimator [13]. Finally, a conservative $Pr(D)$ (Section 3.1) is chosen and the final LR is re-calculated relative to the estimated values of $Pr(D)$, $Pr(C)$ and Fst.

Table 2
Exercise 2 profiles from the victim, the suspect and the unknown sample. Possible allelic drop-out highlighted in grey.

Locus	Victim	Suspect	Unknown sample
D10S1248	14–15	13–14	13–14–15
VWA	15–17	14–17	14–15–17
D16S539	11–11	9–11	9–11
D2S1338	17–24	17–22	17–22
D8S1179	11–13	10–12	10–11–12
D21S11	30–32.2	29–32.2	29–32.2
D18S51	13–14	14–15	13–14–15
D22S1045	15–17	15–17	15–17
D19S433	13–14	13–14	13–14
TH01	6–9.3	6–8	6–8
FGA	20–23	22–23	22–23
D2S441	14–15	11–11	11–15
D3S1358	14–17	16–17	14–16–17
D1S1656	12–15.3	12–16	12–16
D12S391	17–19	17–23	17–23
SE33	14–30.2	28.2–29.2	28.2–29.2

Participants were asked to evaluate the exercises using the following fixed parameters (to prevent unnecessary variation in results):

(a) Set the initial $Pr(D) = 0.5$ to calculate LR before estimating the suitable $Pr(D)$ calculated from the sensitivity analysis where LR is plotted across the range of $Pr(D)$.
(b) Set the drop-in probability $Pr(C) = 0.05$ and Fst = 0 in both exercises.
(c) The same Spanish frequency database was used throughout.

Laboratories were completely free to estimate the number of known and unknown contributors to the crime-samples. From the information provided by the epgs and the case-circumstances, the participants were asked to construct the relevant prosecution (Hp) and defense (Hd) hypotheses. Finally, participants were asked to prepare a statement that described the results for court-going purposes. A total of 18 laboratories participated in this Euroforgen-NoE exercise.

3. Results

A summary of the calculated results is shown in Table 3A (Exercise 1) and Table 3B (Exercise 2).

3.1. Exercise 1 results

All participating laboratories carried out the calculations once except laboratory number 8 who performed 3 independent tests. All of them derived the same pair of hypotheses (column B in Table 3A): (i) Hp: contributors to crime-sample consisted of the victim and the suspect versus (ii) Hd: contributors to the crime-sample consisted of the victim and an unknown random person from the Spanish population. Although Labs 2 and 14 stated the hypotheses taking into account only one contributor, the actual calculations were carried out conditioned on two contributors to the questioned sample (victim and one unknown person).

As expected, the same initial LR was obtained from all participants (4.86×10^{15}; column C in Table 3A) since all of them derived and tested the same hypotheses and used the same initial parameters ($Pr(D) = 0.5$; $Pr(C) = 0.05$; Fst = 0). In order to estimate the plausible range of $Pr(D)$ given the characteristics of the profile in the unknown sample, sensitivity analyses were carried out in LRmix [13]. The analyses provide both the variations of the LR to changes in the drop-out probability in the [0.01, 0.99] interval, and the plausible drop-out values. As the drop-out values are estimated via simulations, slightly different ranges of most plausible $Pr(D)$ were obtained (column F in Table 3A). The LRmix module gave the following ranges of drop-out (taking into account both Hp and Hd): 0.01–0.13; 0.01–0.15; 0.01–0.17 and 0.01–0.19.[1]

The sensitivity test calculates the LR relative to the entire range of $Pr(D)$ (see supplementary Table S3). DNA experts then selected the $Pr(D)$ that minimizes the LR based on 5 or 95 percentiles, and can recalculate the final LR accordingly.

The final LR values reported by the participating laboratories ranged from 6.51×10^{16} to 1.03×10^{17} (Table 3A, column E). All results were within less than one order of magnitude difference.

3.2. Exercise 2 results

Results in this case were more complex since several pairs of hypotheses were taken into account by participating labs; in

[1] Different ranges are obtained because the estimation procedure is based on Monte-Carlo simulations, thus several analyses may lead to slightly different values.

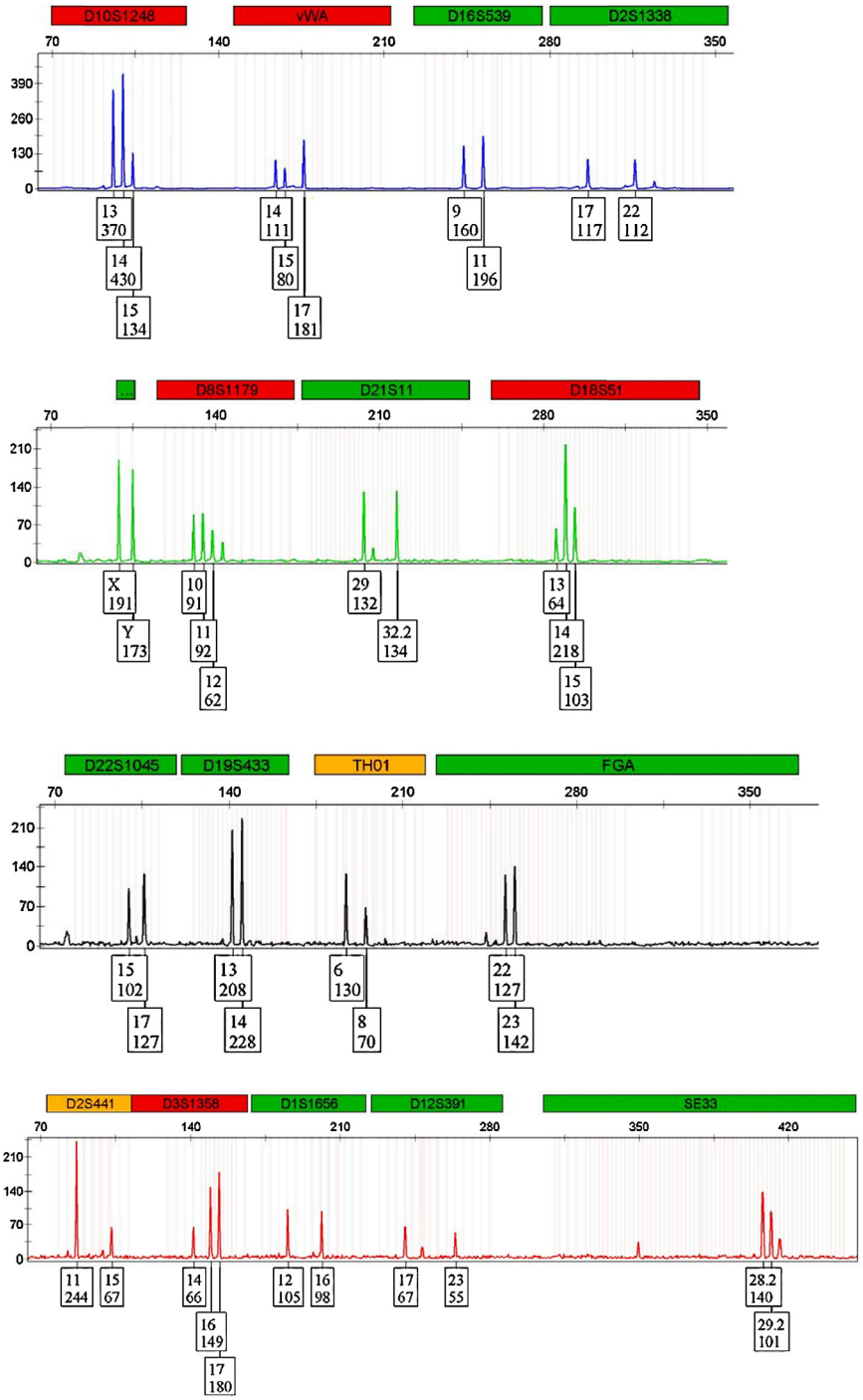

Fig. 1. Electropherogram from the unknown sample in the Exercise 2. Allele calls were performed taking into account a limit of detection of 50 rfus and a true homozygote threshold of 150 rfus.

addition some labs (labs 4, 7 and 18) evaluated the case by using two different pairs of hypotheses and one lab (lab 8) evaluated the same pair twice. In total, 22 evaluations were carried out, as follows:

S = suspect; V = victim; U = unknown

(a) Set 1 propositions: Hp = $S + V$ vs. Hd = $S + U$ (16 evaluations).
(b) Set 2 propositions: Hp = $S + V$ vs. Hd = $U_1 + U_2$ (3 evaluations, all without taking into account the results obtained in locus SE33, see reason below).
(c) Set 3 propositions: Hp = $U + V$ vs. Hd = $U_1 + U_2$ (1 evaluation).
(d) Set 4 propositions: Hp = $S + V$ vs. Hd = $V + U$ (2 evaluations. Note that Lab 14 wrongly stated both hypotheses: Hp = V and Hd = U).

3.2.1. *Set 1 results: Hp* = S + V *vs. Hd* = S + U

All but one laboratory reported that there were two contributors to the questioned sample (suspect and victim/suspect and one unknown person). The only exception was Lab 2, who wrongly stated Hd as a single contributor where only the victim had contributed to the unknown sample, but then proceeded to

Table 3A
Exercise 1 results from 18 participating laboratories (Labs). Laboratory number 8 performed 3 independent tests taking into account the same pair of hypothesis. *Key*: Hp = contributors to the unknown sample under prosecutor hypothesis; Hd = contributors to the unknown sample under the defense hypothesis; LR before sensitivity = likelihood ratio before estimating the drop-out probability given the profile characteristics (an initial *Pr*(*D*) of 0.5 was fixed in order to remove variability in the results); estimated *Pr*(*D*) = probability of drop-out estimated to obtain the lower LR; LR after sensitivity = likelihood ratio after estimating the drop-out probability; *Pr*(*D*) under Hp 5% and 95% percentiles = percentiles 5 and 95 of the distribution of the drop-out probability conditioned on the expected number of alleles observed relative to the genotype of the hypothesized contributors under Hp. *Pr*(*D*) under Hd 5% and 95% percentiles = percentiles 5 and 95 of the distribution of the drop-out probability conditioned on the expected number of alleles observed relative to the genotype of the hypothesized contributors under Hd. Hypotheses wrongly formulated highlighted in grey. Drop-out probabilities wrongly selected to calculate the final LR highlighted in grey.

A	B		C	D	E	F			
Labs	Hp	Hd	LR before sensitivity	Estimated *Pr*(*D*)	LR after sensitivity	*Pr*(*D*) under Hp 5% percentile	*Pr*(*D*) under Hp 95% percentile	*Pr*(*D*) under Hd 5% percentile	*Pr*(*D*) under Hd 95% percentile
1	*V*+*S*	*V*+*U*	4.862×10^{15}	0.19	6.508×10^{16}	0.01	0.11	0.01	0.19
2	*V*+*S*	*V*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.13	0.01	0.15
3	*V*+*S*	*V*+*U*	4.862×10^{15}	0.01	1.032×10^{17}	0.01	0.13	0.01	0.17
4	*V*+*S*	*V*+*U*	4.86×10^{15}	0.17	7.86×10^{16}	0.01	0.11	0.01	0.17
5	*V*+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.13	0.01	0.15
6	*V*+*S*	*V*+*U*	4.86×10^{15}	0.17	7.86×10^{16}	0.01	0.15	0.01	0.17
7	*V*+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.11	0.01	0.15
8	*V*+*S*	*V*+*U*	4.86×10^{15}	0.17	7.86×10^{16}	0.01	0.11	0.01	0.17
	V+*S*	*V*+*U*	4.86×10^{15}	0.01	1.032×10^{17}	0.01	0.13	0.01	0.13
	V+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.11	0.01	0.15
9	*V*+*S*	*V*+*U*	4.862×10^{15}	0.01	1.032×10^{17}	0.01	0.11	0.01	0.17
10	*V*+*S*	*V*+*U*	4.862×10^{15}	0.17	7.857×10^{16}	0.01	0.11	0.01	0.17
11	*V*+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.13	0.01	0.15
12	*V*+*S*	*V*+*U*	4.862×10^{15}	0.17	7.857×10^{16}	0.01	0.13	0.01	0.17
13	*V*+*S*	*V*+*U*	4.86×10^{15}	0.15	9.49×10^{16}	0.01	0.13	0.01	0.15
14	*S*	*U*	4.862×10^{15}	0.01	1.032×10^{17}	0.01	0.09	0.01	0.13
15	*V*+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.11	0.01	0.15
16	*V*+*S*	*V*+*U*	4.862×10^{15}	0.01	1.032×10^{17}	0.01	0.11	0.01	0.13
17	*V*+*S*	*V*+*U*	4.862×10^{15}	0.17	7.857×10^{16}	0.01	0.11	0.01	0.17
18	*V*+*S*	*V*+*U*	4.862×10^{15}	0.15	9.492×10^{16}	0.01	0.13	0.01	0.15

V = victim; S = Suspect; U = Unknown

Table 3B
Exercise 2 results from 18 participating laboratories (Labs). Laboratories number 4 and 7 performed 2 independent tests taking into account 2 different pairs of hypotheses. Laboratory number 8 performed 2 independent tests taking into account the same pair of hypotheses. Key: the same as in Table 3A. Hypotheses wrongly formulated highlighted in grey. Dropout probabilities wrongly selected to calculate the final LR highlighted in grey.

A	B		C	D	E	F			
Lab	Hp	Hd	LR before sensitivity	Estimated *Pr*(*D*)	LR after sensitivity	*Pr*(*D*) under Hp 5% percentile	*Pr*(*D*) under Hp 95% percentile	*Pr*(*D*) under Hd 5% percentile	*Pr*(*D*) under Hd 95% percentile
Set 1 propositions									
1	*S*+*V*	*S*+*U*	7.14×10^{5}	0.19	2.195×10^{5}	0.19	0.43	0.19	0.43
2	*S*+*V*	*S*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.39	0.23	0.43
4	*S*+*V*	*S*+*U*	7.14×10^{5}	0.17	1.52×10^{5}	0.17	0.41	0.19	0.43
6	*S*+*V*	*S*+*U*	7.14×10^{5}	0.21	2.93×10^{5}	0.21	0.43	0.21	0.45
7	*S*+*V*	*S*+*U*	7.145×10^{5}	0.17	1.52×10^{5}	0.17	0.43	0.19	0.45
8	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.43	0.19	0.43
8	*S*+*V*	*S*+*U*	7.145×10^{5}	0.17	1.52×10^{5}	0.17	0.43	0.21	0.45
9	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.43	0.21	0.45
10	*S*+*V*	*S*+*U*	1.178×10^{6}[a]	0.15	3.634×10^{5}[a]	0.15	0.41	0.19	0.45
11	*S*+*V*	*S*+*U*	7.145×10^{5}	0.17	1.52×10^{5}	0.17	0.43	0.25	0.45
12	*S*+*V*	*S*+*U*	7.145×10^{5}	0.21	2.93×10^{5}	0.21	0.43	0.21	0.47
13	*S*+*V*	*S*+*U*	7.145×10^{5}	0.21	2.93×10^{5}	0.21	0.43	0.21	0.47
15	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.41	0.19	0.45
16	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.43	0.21	0.43
17	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.41	0.21	0.47
18	*S*+*V*	*S*+*U*	7.145×10^{5}	0.19	2.195×10^{5}	0.19	0.43	0.19	0.43
Set 2 propositions									
4	*S*+*V*	*U*+*U*	8.99×10^{16}[a]	0.43	1.6×10^{17}[a]	0.19	0.41	0.19	0.43
5	*S*+*V*	*U*+*U*	8.992×10^{16}[a]	0.45	1.359×10^{17}[a]	0.15	0.39	0.17	0.45
18	*S*+*V*	*U*+*U*	8.992×10^{16}[a]	0.41	1.89×10^{17}[a]	0.17	0.41	0.19	0.41
Set 3 propositions									
7	*U*+*V*	*U*+*U*	5759	0.21	1091	0.21	0.49	0.21	0.43
Set 4 propositions									
14	*V*	*U*	2.083×10^{14}	0.42	4.683×10^{14}	0.19	0.41	0.21	0.45
3	*S*+*V*	*V*+*U*	2.083×10^{14}	0.21	6.897×10^{15}	0.19	0.43	0.21	0.45

V = victim; *S* = suspect; *U* = unknown.
[a] LRs without SE33 marker.

carry out calculations correctly conditioned on two ($V + U$) contributors (Supplementary Table S5).

As expected, in all but one of the 16 evaluations made, the same initial LR was obtained (7.14×10^5; column C in Table 3B) since all tested the same hypotheses and used the same initial parameters ($Pr(D) = 0.5$; $Pr(C) = 0.05$; Fst = 0). Lab 10 obtained a different result in this initial LR (1.18×10^6) because the results at SE33 locus were not taken into account (the lab reported computer software/memory problems with this marker with LRmix, since resolved – see solution in Section 3.2.2).

Regarding the sensitivity test, and taking into account the results from the 15 evaluations and both Hp and Hd, the widest (5–95 percentile) range of $Pr(D)$ was 0.17–0.47 (column F in Table 3B and Supplementary Table S4A). In the case of Lab 10 this range was 0.15–0.45. In all cases the estimated $Pr(D)$ used in the calculation was the one giving the lowest LR. Removing the Lab 10 data from the summary (as SE33 had been omitted), three different final LR values were obtained: 2.93×10^5 (Pr(D) = 0.21), 2.19×10^5 (Pr(D) = 0.19) and 1.52×10^5 ($Pr(D) = 0.17$) (Table 3B, column E). The variation in results was minor, and all were of the same order of magnitude (10^5) and are therefore consistent.

3.2.2. Set 2 results: *Hp* = S + V *vs.* *Hd* = $U_1 + U_2$

The three evaluations carried out with these hypotheses did not take into account the electropherogram data obtained in the SE33 locus. The exercise took place in November–February 2012–2013 when Forensim v 3.2.1 was available. This version needed rapid computers with high capacity to analyze numerous allelic combinations. Not all computers available to participants had the capacity to process such a large amount of information. The current version of Forensim (v 4.0) programmed in C considerably speeds up the calculations and has rectified this problem.

The three labs obtained the same initial LR (8.99×10^{16}; column C in Table 3B). Following the sensitivity test and taking into account the results from the 3 evaluations, the widest range of $Pr(D)$ was 0.15–0.45 (column F in Table 3B and Supplementary Table S4B). Three similar final LR values were obtained: 1.89×10^{17} ($Pr(D) = 0.41$), 1.6×10^{17} ($Pr(D) = 0.43$) and 1.36×10^{17} ($Pr(D) = 0.45$) (Table 3B, column E). Once again, consistency of results was demonstrated.

3.2.3. Set 3 results: *Hp* = U + V *vs.* *Hd* = $U_1 + U_2$

Only Lab 7 took into account this pair of hypotheses where the profile of the suspect was not considered as a contributor to the evidence under either hypothesis. The initial LR before performing the sensitivity test was 5759 (column C in Table 3B). The plausible range of $Pr(D)$ was 0.21–0.49 (Supplementary Table S4C) and the final LR (applying $Pr(D) = 0.21$) was 1091 (Table 3B, column E).

3.2.4. Set 4 results: *Hp* = S + V *vs.* *Hd* = V + U

Recall that the judge in this case requested the crime-sample to be evaluated to discover if there was any evidence for the victim's DNA present on the knife (see Materials and Methods). Therefore the defence hypothesis Hd = V was not appropriate for this purpose. Labs number 3 and 14 carried out the case evaluation using this pair of alternative hypotheses. Lab 14 also stated both hypotheses as: Hp = V vs. Hd = U, but the actual calculations performed were: Hp = $S + V$ vs. Hd = $V + U$. Both labs obtained the same LR before the sensitivity test (2.08×10^{14}, column C in Table 3B) and both labs applied a non-conservative $Pr(D)$ to calculate the final LR after the sensitivity test (Supplementary Table S4D).

4. Statements

In both Exercises 1 and 2, all but one laboratory (Lab 14) defined their conclusions with regard to the calculated LR. Different expressions were used by the participants (see Supplementary Tables S5 and S6). The statements perfectly reflected the meaning of the LR except in the case of two labs (2 and 3) who reported their conclusions by transposing the conditional (LR = $Pr(\text{Hp}|E)/Pr(\text{Hd}|E)$ instead of LR = $Pr(E|\text{Hp})/Pr(E|\text{Hd})$).

5. Discussion

To the best of our knowledge this is the first report on a collaborative exercise on statistical evaluation of complex mixtures using a probabilistic method. A previous unpublished exercise (MIX05) reported by NIST showed that laboratories were inconsistent in their reporting [17].

In this exercise, a total of 18 laboratories participated in the study, where two hypothetical case examples of DNA mixed profiles were evaluated. Both profiles presented one major and one minor contributor where the major was a complete profile. The first mixed profile (Exercise 1) perfectly matched the reference profiles except at one marker (SE33) where a drop-out event occurred (allele 24.2 from the questioned contributor, the suspect). The second one (Exercise 2) showed ten drop-out events (allele 24 in D2S1338; allele 13 in D8S1179; allele 30 in D21S1; allele 9.3 in TH01; allele 20 in FGA; allele 14 in D2S441; allele 15.3 in D1S1656; allele 19 in D12S391 and alleles 14 and 30.2 in SE33), all of them from the questioned contributor (victim).

The aim of the exercise was to test if standardization could be achieved by using the LRmix program. Participants were supplied with the crime-sample and the reference DNA profiles. The same population database of allele frequencies was provided to the participants in order to avoid variation in LR values due to different allelic frequencies.

5.1. Exercise 1

In Exercise 1 the files containing the genotypes from evidence, victim and suspect were sent to the participants. These were used to generate the profiles uploaded to LRmix. Although no guidelines were supplied to the laboratories, all the participants tested the same pair of hypotheses (Hp = $V + S$ vs. Hd = $V + U$). The final LRs ranged from 6.51×10^{16} to 1.03×10^{17} which means that high uniformity was obtained among the laboratories. The slight variation is due to the chosen value of drop-out probability from the plausible range of values estimated by the program. A more detailed explanation about how $Pr(D)$ in LRmix is calculated can be found in Supplementary material S7.

5.2. Exercise 2

Exercise 2 differed from Exercise 1 in that participants were free to choose alleles for analysis. Two files containing the reference sample profiles were provided, while the evidence profile was given as an electropherogram (with filtered stutters following the recommendations of the kit manufacturer and a limit of detection of 50 rfu). No participants included signals under 50 rfu despite some of these clearly exceeding the background signal (see Fig. 1, allele 13 in D8S1179 as an example). In earlier inter-laboratory exercises [18] it was demonstrated that participants could correctly identify all major-donor alleles in mixed samples but the interpretation of minor-donor alleles varied significantly between laboratories.

In the absence of probabilistic approaches, analysts are forced to make subjective binary decisions about below-threshold alleles that are consequently designated as a true homozygote or as a heterozygote (with dropout). The combination of probabilistic methods combined with probabilistic methods of interpretation substantially reduces the need to make arbitrary subjective

decisions. This in turn promotes standardization, by driving forward a uniformity of approach.

The majority of participants evaluated Exercise 2 under the "Set 1 Propositions" (Hp = $S + V$ vs. Hd = $S + U$). This was a logical approach since the suspect had admitted that he had cut his hand with the knife, but had denied murder, consequently the judge had asked if there was evidence to suggest that the victim was also a contributor. The final LRs reported by the participants showed only slight differences, ranging from 1.52×10^5 to 2.93×10^5. Again, this minor deviation was due to the estimated probabilities of drop-out.

Three participants evaluated the "Set 2 Propositions" (Hp = $S + V$ vs. Hd = $U_1 + U_2$). This was also a valid set of propositions. In this case, the judge only mentioned the presence or absence of the victim in the sample, so the question of additional unknown contributors was a matter for the expert to evaluate. One of the aims of the Exercise 2 was to determine how laboratories deal with propositions that hypothesize different numbers of contributors. Indeed, some laboratories carried out the evaluation twice by using the former and the present pair of hypotheses. Although only three participants evaluated this set of hypotheses, the obtained values only ranged from 1.36×10^{17} to 1.89×10^{17}. This LR was much higher than that obtained with the first pair of hypotheses (LR = 3.63×10^5, results of Lab 10 who carried out the calculations removing the SE33 results). This demonstrated that applying an additional unknown contributor to Hd did not act in favour of the defense hypothesis and was the most conservative estimate.

One participant evaluated the "Set 3 Propositions" (Hp = $V + U$ vs. Hd = $U_1 + U_2$). In this case neither the statement information nor the genetic information from the suspect was taken into account. Comparing the value (LR = 1091) with the range of LRs of the first (10^5) or second group (10^{17} without SE33) it was shown that very different final LRs were obtained when comparing different pairs of hypotheses, but there was very marked consistency whenever the same hypotheses were evaluated. This also highlighted the exploratory nature of the method [13] where the evaluation of different sets of hypotheses is to be strongly encouraged. An important part of any training course is to provide guidelines and examples of how to formulate the propositions. It will be often necessary for the expert to evaluate different sets of propositions – often there is not just a single pair to consider. For this reason, probabilistic methods must be considered as part of a holistic approach that incorporates a meaningful evaluation of a case in its entirety. The 'black box' approach is strongly discouraged; more than one answer is possible.

Finally, two laboratories tested a pair of hypothesis that did not answer the question formulated by the judge since both considered the victim was a true contributor to the mixture under Hd. This is not considered further here.

5.3. Reporting the LR meaning(statement) to the court

Laboratory 14 did not explain the LR results in any of the exercises. This lab also misunderstood the sensitivity test and tested hypotheses that did not answer the question formulated by the judge in Exercise 2. But most of laboratories (all but two) correctly reported the conclusion that can be reached from the statistical evaluation in both exercises. The two laboratories (2 and 3) that wrongly reported the conclusions made the same mistake – transposition of the conditional – which is the most frequent error made when writing LR statements. It should be mentioned that these labs also made other mistakes. Laboratory 2 wrongly formulated the Hd hypothesis in both exercises and Laboratory 3 made the same mistakes as Laboratory 14 (misunderstood the sensitivity test and tested hypotheses that did not answer the question formulated by the judge in the Exercise 2). Therefore, all deviations were observed in a subset of 3 laboratories.

6. Conclusions

Undoubtedly the Euroforgen-NoE educational effort to organize a course to train DNA experts was one of the keys of the successful results obtained in the present collaborative exercise. The probabilistic approach to mixture interpretation reduces subjectivity in expert opinion, since uncertainty in the DNA profile, for example ambiguous alleles, can be accounted for in the likelihood ratio. The exercise illustrates that there is no single answer, since there are always unknown aspects to any case (e.g. the number of contributors) but a format can be followed to evaluate evidence using the exploratory approach offered by LRmix [8].

Interpretation of mixtures and/or low level DNA profiles has proven to be one of the most complex tasks in the forensic field. Classical LR approaches used a binary model that forced decisions to be made about reporting peak signals near stochastic thresholds, since the numerator only could take 0 or 1 values (binary approach: exclusion/inclusion criteria). The new semi-continuous LR theory enables an increase in the number of complex cases that can be evaluated and reported to the court but LR formulation becomes more complex and the use of software is required to carry out these types of evaluations.

In this inter-laboratory exercise we have demonstrated that the LRmix program within Forensim is a useful tool to deal with complex scenarios and that a high level of standardization is obtained between laboratories. The allele designations of a complex mixture were reproducible and no major deviations were detected among laboratories. Secondly, we have demonstrated that several scenarios can be rapidly evaluated with LRmix (different pairs of hypotheses were tested by laboratories). Finally, similar statistical results were obtained by the participating laboratories, which is highly desirable from the Court perspective to demonstrate consistency.

In conclusion, in this paper we have demonstrated that the standardization of the probabilistic evaluation is possible, provided the same sets of hypotheses are compared, when suitable tools and training is provided to the DNA forensic experts.

Acknowledgements

The authors express their sincere gratitude to Carlos del Valle for his help in the managing of the Euroforgen-NoE course. Peter Gill has received funding support from the European Union Seventh Framework Programme (FP7/2007–2013) under Grant agreement no. 285487 (Euroforgen-NoE). The input of Hinda Haned was partly supported by a grant from the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.10.011.

References

[1] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.

[2] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.

[3] P. Gill, R. Puch-Solis, J. Curran, The low-template-DNA (stochastic) threshold-its determination relative to risk analysis for national DNA databases, Forensic Sci. Int. Genet. 3 (2009) 104–111.
[4] H. Haned, T. Egeland, D. Pontier, L. Pene, P. Gill, Estimating drop-out probabilities in forensic DNA samples: a simulation approach to evaluate different models, Forensic Sci. Int. Genet. 5 (2011) 525–531.
[5] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, M. Prinz, T. Caragine, Likelihood ratio statistics for DNA mixtures allowing for drop-out and drop-in, Forensic Sci. Int. Genet. (Suppl. Ser.) 3 (2011) e240–e241.
[6] T. Tvedebrink, P.S. Eriksen, M. Asplund, H.S. Mogensen, N. Morling, Allelic dropout probabilities estimated by logistic regression—further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2012) 263–267.
[7] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (2009) e8327.
[8] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[9] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.
[10] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules f or STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[11] C.C. Benschop, C.P. van der Beek, H.C. Meiland, A.G. van Gorp, A.A. Westen, T. Sijen, Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results, Forensic Sci. Int. Genet. 5 (2011) 316–328.
[12] C.C. Benschop, H. Haned, T.J. de Blaeij, A.J. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: A descriptive study, Forensic Sci. Int. Genet. 6 (2012) 697–707.
[13] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (2012) 762–774.
[14] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, J. Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Sci. Int. Genet. 2 (2008) 91–103.
[15] J. Curran, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.
[16] http://www.cstl.nist.gov/strbase/interlab/MIX05/MIX05poster.pdf, march 2013.
[17] http://www.gep-isfg.org/es/comisiones-trabajo/ejercicio-colaborativo-ghep-mix-2009/resultados-ejercicio-colaborativo-sobre-analisis.html, march 2013.
[18] D.L. Duewer, M.C. Kliner, J.W. Redman, J.M. Butler, NIST mixed stain study 3: signal intensity balance in commercial short tandem repeat multiplexes, Anal. Chem. 76 (2004) 6928–6934.
[19] D.J. Balding, Weight-of-Evidence for Forensic DNA Profiles, John Wiley and Sons Ltd., 2005.